VALIDADE DE TESTES

Sérgio Fukusima

Depto Psicologia / FFCLRP

DEFINIÇÃO DE VALIDADE

“A validade de um teste diz respeito a o que o teste mede e com que eficiência ele faz.”

(Anastasi & Urbina)

Implicação:

1. a validade é uma propriedade dos testes, e não das interpretações de seus escores;
2. para serem válidos, os escores de teste devem medir algum suposto constructo diretamente;
3. a validade de um escore é, pelo menos em certo grau, uma função da compreensão do autor ou desenvolvedor do teste a respeito do constructo que ele pretende medir.

Não esqueça

As pessoas fazem inferências a partir de observações e amostras de comportamento o tempo todo. Por exemplo, se escutamos alguém falar com muitos erros gramaticais, podemos inferir que esta pessoa tem baixo nível de escolaridade. Se uma pessoa chega invariavelmente na hora marcada, podemos inferir que ela é pontual. Algumas de nossas inferências são corretas, e algumas não. Algumas são importantes, e outras não.

Se as inferências que fazemos são importantes o bastante para desejarmos determinar sua correção, ou seja, validá-las, precisamos

1. definir nossos termos inequivocamente (p. ex., o que queremos dizer com "escolaridade"? "Chegar sempre na hora marcada" representa plenamente o conceito de pontualidade?);
2. investigar a fidedignidade de nossas observações (p. ex., a pessoa sempre comete erros gramaticais, ou apenas em algumas circunstâncias? Nosso amigo chega na hora marcada em todos os seus compromissos, ou apenas naqueles que tivemos oportunidade de observar?);
3. decidir se existem evidências suficientes para justificar as inferências que queremos fazer com base em nossas definições e nos dados disponíveis (p. ex., chegar na hora marcada em todos os compromissos é base suficiente para se julgar a pontualidade de uma pessoa), ou se precisamos corroborar nossas inferências com mais dados (p. ex., a pessoa demonstra outros indicadores daquilo que queremos dizer com "baixo nível de escolaridade"?).

Os testes psicológicos são ferramentas criadas para ajudar a refinar e quantificar observações comportamentais para fins de interferências a respeito de indivíduos, grupos ou constructos psicológicos. Fundamentalmente, os escores de testes psicológicos são válidos se podem nos ajudar a fazer inferências precisas.

Tabela 5.1 Aspectos da validade de constructo e fontes de evidências relacionadas

Aspecto da validade do constructo	Fontes de evidências[a]
Relacionada ao conteúdo	Relevância e representatividade do conteúdo do teste e dos processos de resposta às tarefas Validade de face (isto é, aparência superficial)
Padrões de convergência e divergência	Consistência interna de resultados do teste e outras medidas de fidedignidade Correlações entre testes e subtestes Matriz multitraço-multimétodo Diferenciação de escores de acordo com diferenças esperadas com base na idade e outras variáveis de *status* Resultados experimentais (isto é, correspondência entre escores de teste e os efeitos preditos de intervenções experimentais ou hipóteses baseadas em teorias) Análise fatorial exploratória Técnicas de modelagem de equação estrutural
Relacionada ao critério	Precisão das decisões baseadas na validação concorrente (isto é, correlações entre escores de teste e critérios existentes) Precisão de decisões ou predições baseadas na validação preditiva (isto é, correlações entre escores de testes e critérios preditos)

[a] Ver Capítulo 5 para explicações dos termos.

CONSULTA RÁPIDA 5.2

Desconstruindo constructos

Como o termo *constructo* é usado com tanta freqüência neste capítulo, um esclarecimento do seu sentido é necessário. De modo geral, um *constructo* é qualquer coisa criada pela mente humana que não seja diretamente observável. Os constructos são abstrações que podem se referir a conceitos, idéias, entidades teóricas, hipóteses ou invenções de muitos tipos.

Na psicologia, o termo *constructo* é aplicado a conceitos como traços, e às relações teóricas entre conceitos que são inferidas de observações empíricas consistentes de dados comportamentais. Os constructos psicológicos diferem amplamente em termos de

- sua amplitude e complexidade,
- sua aplicabilidade potencial e
- grau de abstração necessário para inferi-los a partir dos dados disponíveis.

Como regra, constructos de definição estrita requerem menos abstração, mas têm uma gama menor de aplicações. Além disso, como é mais fácil obter consenso a respeito de constructos estritos, simples e menos abstratos, estes também são avaliados com mais facilidade do que constructos mais amplos e multifacetados que podem ter adquirido sentidos diferentes em vários contextos, culturas e períodos históricos.

Exemplos:

- Enquanto a *destreza manual* é um constructo que pode ser relacionado prontamente a dados comportamentais específicos, a *criatividade* é muito mais abstrata. Por isso, quando é necessário avaliar esses traços, determinar quem tem mais destreza manual é muito mais fácil do que determinar quem é mais criativo.
- A *introversão* é um constructo mais simples e de definição mais estrita do que a *conscienciosidade*. Embora esta seja potencialmente útil na predição de uma gama mais ampla de comportamentos, ela também é mais difícil de avaliar.

Sinônimos: os termos *constructo* e *variável latente* muitas vezes são usados de forma equivalente. Uma *variável latente* é uma característica que presumivelmente subjaz a um fenômeno observado, mas não é diretamente mensurável ou observável. Todos os traços psicológicos são variáveis latentes, ou constructos, assim como as denominações dadas a fatores que emergem de pesquisas de análise fatorial, como compreensão verbal ou neuroticismo.

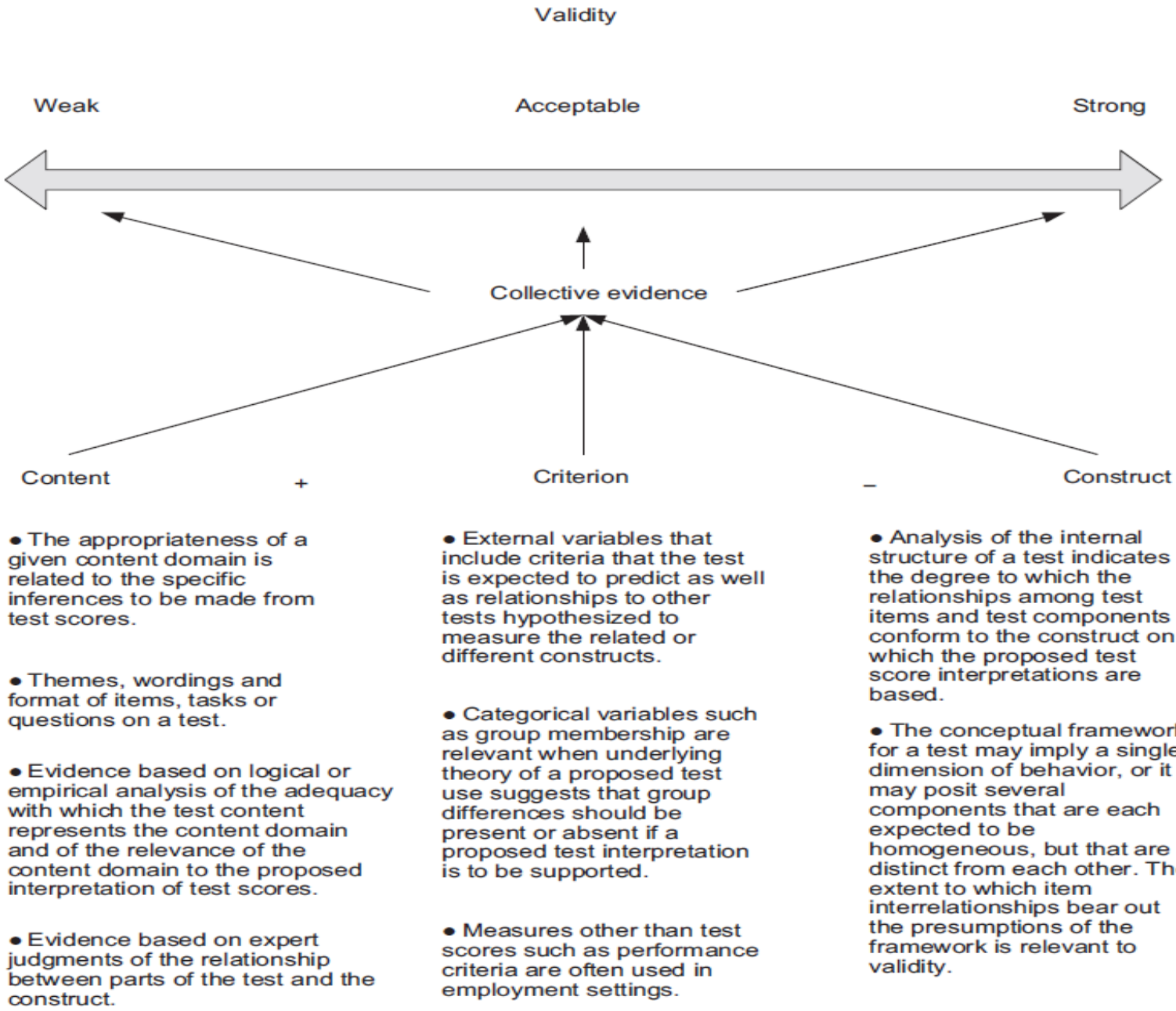

FIGURE 3.1. Validity continuum. Bulleted information is from AERA, APA, and NCME (1999, pp. 11–13).

FONTES DE EVIDÊNCIAS DE VALIDADE

Evidências de validade baseadas no conteúdo do teste e processos de resposta

Testes Educacionais
Testes Ocupacionais

CONSULTA RÁPIDA 5.3

Exemplos de testes educacionais padronizados que usam evidências baseadas no conteúdo como principal fonte de validação

Título do teste	Objetivo principal	Aplicações primárias	Site da Internet com descrição e amostras do teste
Test of English as a Foreign Language (TOEFL)	Avaliar a proficiência em inglês de pessoas cuja língua nativa não seja o inglês	Determinar se estudantes estrangeiros possuem conhecimento suficiente de inglês para serem admitidos em faculdades americanas	http://www.toefl.org
College-Level Examination Program (CLEP) Introductory Psychology Test	Medir o conhecimento dos materiais habitualmente ensinados em disciplina introdutória de psicologia em um semestre	Determinar se os estudantes têm conhecimento suficiente de psicologia introdutória para receber um crédito universitário através de exame	http://www.collegeboard.com/clep
ACT Assessment Science Reasoning Test	Medir as habilidades de interpretação, análise, avaliação, raciocínio e solução de problemas necessárias no campo das ciências naturais, incluindo biologia, química, física e ciências espaciais	Avaliar o conhecimento e as habilidades adquiridas por um estudante para determinar sua capacidade para assumir empregos de nível universitário	http://www.act.org
National Assessment of Educational Progress (NAEP)	Medir conhecimentos e habilidades em leitura, matemática, ciências, escrita, história dos EUA, geografia e artes	Fornecer informações a respeito do desempenho de populações e subgrupos de estudantes em todos os EUA e estados participantes	http://nces.ed.gov

CONSULTA RÁPIDA 5.4

Exemplos de testes ocupacionais padronizados que usam evidências baseadas no conteúdo como fonte de validação

Título do teste	Constructo avaliado	Descrição	Aplicação primária
Crawford Small Parts Dexterity Test (CSPDT)[a]	Coordenação visual-manual e destreza motora fina	O CSPDT consiste em duas tarefas: (a) trabalhar com pinças inserindo pequenos alfinetes nos orifícios de uma bandeja e depois colocar pequenos aros sobre as partes projetadas dos alfinetes; (b) inserir parafusos na bandeja e depois apertá-los com uma chave de fenda. A velocidade do desempenho é o principal fator na avaliação deste teste.	Usado para determinar se um indivíduo tem a destreza manual necessária para qualquer emprego que envolva trabalho de precisão com as mãos, como entalhes ou conserto de relógios.
Clerical Abilities Battery (CAB)[a]	Diversos componentes de uma ampla gama de ocupações administrativas identificadas pela análise de função de comportamentos administrativos gerais	O CAB tem sete subtestes auto-explicativos: Arquivamento, Comparação de Informações, Cópia de Informações, Uso de Tabelas, Revisão, Habilidades Básicas em Matemática e Raciocínio Numérico.	Usado para recrutamento e avaliação de funcionários administrativos. Em uma revisão do *Mental Measurements Yearbook*, Randhawa (1992) afirma que a amostragem e a amplitude das tarefas dos subtestes do CAB não são suficientemente representativos, e sugere que são necessários mais dados de padronização, fidedignidade e validade preditiva. No entanto, ele admite que o processo de desenvolvimento e o formato da bateria são adequados e fornecem as bases para uma ferramenta potencialmente excelente.

[a]Publicado por Psychological Corporation (http://www.PsychCorp.com).

Evidências de validade de conteúdo em outros contextos de avaliação

Evidências de validade do ponto de vista dos testandos

Evidências de validade baseadas na exploração de padrões de convergência e divergência

A fidedignidade dos escores como fonte de evidência de validade

Correlações entre testes e subtestes

Validade de Critério: Correlações dos escores de testes com critérios externos

- **Validade concorrente**
- **Validade preditiva**

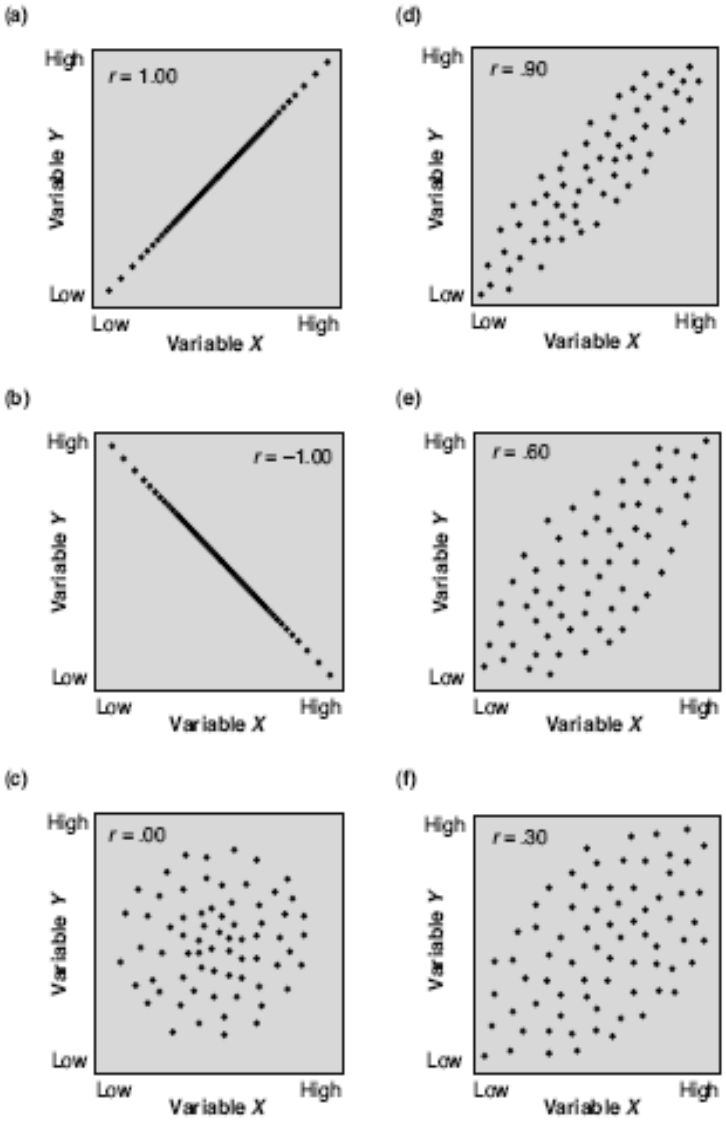

FIGURE 9 Scatterplot of Different Correlation Coefficients.

Source: Hopkins, Kenneth, *Educational and Psychological Measurement and Evalutation*, 8th ©1998. Printed and Electronically reproduced by permission of Pearson Education, Inc., Upper Saddle River, New Jersey.

TABLE 6 Calculating a Pearson Correlation Coefficient

There are different formulas for calculating a Pearson correlation coefficient and we will illustrate one of the simpler ones. For this illustration we will use the test scores we have used before as the *X* variable, and another set of 20 hypothetical scores as the *Y* variable. The formula is:

$$r_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma X)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma X)^2}}$$

XY = sum of the XY products
X = sum of X scores
Y = sum of Y scores
X^2 = sum of squared X scores
Y^2 = sum of squared Y scores

Test 1 (*X*)	X^2	Test 2 (*Y*)	Y^2	(*X*)(*Y*)
7	49	8	64	56
8	64	7	49	56
9	81	10	100	90
6	36	5	25	30
7	49	7	49	49
6	36	6	36	36
10	199	9	81	90
8	64	8	64	64
5	25	5	25	25
9	81	9	81	81
9	81	8	64	72
9	81	7	49	63
8	64	7	49	56
4	16	4	16	16
5	25	6	36	30
6	36	7	49	42
7	49	7	49	49
8	64	9	81	72
8	64	8	64	64
7	49	6	36	42
X = 146	X^2 = 1,114	Y = 143	Y^2 = 1,067	XY = 1,083

$$r_{xy} = \frac{20(1{,}083) - (140)(143)}{\sqrt{20(1{,}114) - (146)^2}\sqrt{20(1{,}067) - (143)^2}}$$

$$\frac{21{,}660 - 20{,}878}{\sqrt{22{,}280 - 21{,}316}\sqrt{21{,}340 - 20{,}449}} = \frac{782}{\sqrt{964}\sqrt{891}}$$

$$\frac{782}{(31.048)(29.849)} = 0.843$$

Validade Preditiva

Tabela 5.4 Dados para o exemplo de validação preditiva

Candidato	Escore no teste (X)	Produção (Y)	$X - M_x$ (x)	$Y - M_y$ (y)	x^2	y^2	xy
1	18	56	5	6	25	36	30
2	12	50	-1	0	1	0	0
3	8	47	-5	-3	25	9	15
4	20	52	7	2	49	4	14
5	14	52	1	2	1	4	2
6	5	42	-8	-8	64	64	64
7	10	48	-3	-2	9	4	6
8	12	49	-1	-1	1	1	1
9	16	50	3	0	9	0	0
10	15	54	2	4	4	16	8
Soma130	500	0	0	188	138	140	
Média13	50						

Tabela 5.5 Análise dos dados de validação preditiva

Estatística descritiva	Número de observações (N)	Média	Desvio padrão (DP)
X = preditor (escores de teste)	10	13	4,57
Y = critério (produção)	10	50	3,91

Pearson r $$r_{xy} = \frac{\sum xy}{(N-1)(DPx)(DPy)} = \frac{140}{(9)(4,57)(3,91)} = 0,87$$

Coeficiente de determinação $r_{xy}^2 = 0,755$

Equação de regressão linear $Y' = a_{yx} + b_{yx}(X)$

Dados do exemplo $a_{yx} = 40,32$
$b_{yx} = 0,745$

Equação regressiva para predizer a produção da linha de montagem com base nos escores do teste de destreza

Produção predita = $Y' = 40,32 + 0,745(X)$

*Onde Y' = escore predito no critério; $a_{yx} = M_y - b_{yx}(M_x)$ = intercepto da linha de regressão; $b_{yx} = (\sum xy)/(\sum x^2)$ = declividade da linha de regressão; e X = escore no preditor [escore no teste de destreza manual].

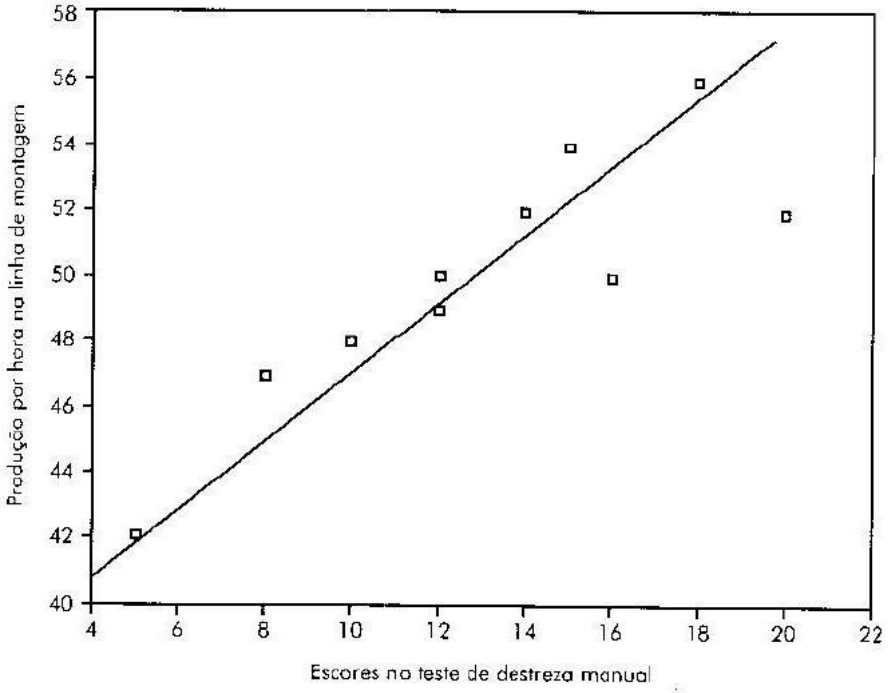

Tabela 5.2 Uma matriz multitraços-multimétodos hipotética (MTMMM)

		Auto-Relato			Observação			Projetiva		
Método	Traço	Ans	Afi	Dom	Ans	Afi	Dom	Ans	Afi	Dom
Auto-Relato	Ans	(0,90)								
	Afi	*0,45*	(0,88)							
	Dom	*0,35*	*0,38*	(0,80)						
Observação	Ans	**0,60**	0,23	0,10	(0,92)					
	Afi	0,25	**0,58**	-0,08	*0,47*	(0,93)				
	Dom	0,12	-0,12	**0,55**	*0,30*	*0,32*	(0,86)			
Projetiva	Ans	**0,56**	0,22	0,11	**0,65**	0,40	0,31	(0,94)		
	Afi	0,23	**0,57**	0,05	0,38	**0,70**	0,29	*0,44*	(0,89)	
	Dom	0,13	-0,10	**0,53**	0,19	0,26	**0,68**	*0,40*	*0,44*	(0,86)

Nota: Ans = ansiedade; Afi = afiliação; Dom = dominância. Os coeficientes de fidedignidade estão entre parênteses, ao longo da diagonal principal. Os coeficientes de validade (mesmo traço avaliado por diferentes métodos) estão em negrito. Todos os outros coeficientes são índices da validade discriminante de escores de traços diferentes avaliados por um único método (representando a variância com o mesmo método em itálico) e traços diferentes avaliados por métodos diferentes (letra simples).

Análise Fatorial

Tabela 5.3

A. Matriz de correlação: intercorrelações de escores em cinco subtestes do Beta III para 95 estudantes universitários

Subteste	Codificação	Completar desenhos	Checagem administrativa	Absurdos em desenhos	Raciocínio matricial
Codificação	1,00	0,13	0,62**	0,20	0,05
Completar desenhos		1,00	0,09	0,21*	0,11
Checagem administrativa			1,00	0,18	0,20
Absurdos em desenhos				1,00	0,31**
Raciocínio matricial					1,00

Nota: Os dados são de Urbina e Ringby (2001).
*p= .05 **p= .01

B. Matriz fatorial para os dois fatores extraídos da análise fatorial exploratória (AFE) de cinco subtestes do Beta III[a]

Subteste	Cargas no fator 1	Cargas no fator 2
Codificação	**0,90**	0,06
Completar desenhos	0,07	**0,54**
Checagem administrativa	**0,88**	0,14
Absurdos em desenhos	0,15	**0,75**
Raciocínio matricial	0,02	**0,73**

Nota: Os números em negrito indicam as cargas mais altas nos dois fatores transformados por rotação varimax.
[a]Os fatores 1 e 2 respondem por 61% da variância nos escores dos subtestes; os 39% restantes da variância são explicados por fatores específicos de cada subteste e pela variância de erro.

Estratégias de validação em relação à interpretação de escores de teste

Teste cujos escores serão interpretados	Objetivo proposto da interpretação dos escores do teste	Tipo de estratégia de validação desejada	Possíveis fontes de evidências
Exame final da disciplina Cálculo I	Determinar se os estudantes serão aprovados na disciplina Cálculo I	Conteúdo	Relevância e representatividade do conteúdo do teste em relação aos temas abordados na disciplina Cálculo I
	Determinar se os estudantes estão prontos para a disciplina Cálculo II	Relacionada ao critério, tipo concorrente	Correlação positiva alta entre escores no teste de Cálculo I e notas na disciplina Cálculo II
	Predizer se os estudantes podem completar com sucesso a graduação em matemática	Relacionada ao critério, tipo preditivo	Correlação positiva alta entre escores no teste de Cálculo I e conclusão do curso de matemática
	Investigar a relação entre a habilidade matemática e o tipo de personalidade	Convergência	Suporte para a hipótese de que estudantes introvertidos vão ter escore mais alto do que estudantes extrovertidos no teste de Cálculo I